




Hierarchical Clustering



Herarchical Clustering

- ▶ Se le llama “clustering” a la formación de grupos o agrupaciones entre muestras o entre variables.
 - ▶ Los “clusters” son formados por datos que cumplen una condición de agrupación definida en función de algún índice de similitud.
 - ▶ Los agrupamientos no siempre serán los mismos, ya que depende del tipo de dato y también depende del tipo de índice de similitud.
 - ▶ El “Herarchical Clustering” es un algoritmo exhaustivo que realiza la comparación de todos los datos contra todos. Además es clasificado como un algoritmo no supervisado, porque no se conoce de antemano los resultados del proceso de “clustering”.
- 

Herarchical Clustering

El proceso de "clustering" comienza con el cálculo de las similitudes entre muestras o entre variables.

	A	B	C	D	E	F	G
A	0	0.5	0.43	1	0.25	0.63	0.38
B	0.5	0	0.71	0.83	0.67	0.20	0.78
C	0.43	0.71	0	1	0.43	0.67	0.33
D	1	0.83	1	0	1	0.80	0.86
E	0.25	0.67	0.43	1	0	0.78	0.38
F	0.63	0.20	0.67	0.80	0.78	0	0.75
G	0.38	0.78	0.33	0.86	0.38	0.75	0


Herarchical Clustering

- Como es casi obvio, el primer paso es decidir el par de muestras o variables que según el índice de similitud sean los más cercanos.

	A	B	C	D	E	F	G
A	0	0.5	0.43	1	0.25	0.63	0.38
B	0.5	0	0.71	0.83	0.67	0.20	0.78
C	0.43	0.71	0	1	0.43	0.67	0.33
D	1	0.83	1	0	1	0.80	0.86
E	0.25	0.67	0.43	1	0	0.78	0.38
F	0.63	0.20	0.67	0.80	0.78	0	0.75
G	0.38	0.78	0.33	0.86	0.38	0.75	0




Herarchical Clustering

- ▶ Las variables más similares son usadas para crear un cluster o grupo.
 - ▶ Esto equivale a crear una nueva muestra o nueva variable combinando las similitudes de los pares más similares.
 - ▶ El criterio más usado es el llamado "complete linkage".
 - ▶ Este criterio se basa en crear una nueva muestra o variables eligiendo la máxima similitud del par similar contra los demás datos.
- 

Herarchical Clustering

- Las variables más similares son usadas para crear un cluster o grupo.
- Esto equivale a crear una nueva muestra o nueva variable combinando las similitudes de los pares más similares.
- El criterio más usado es el llamado "complete linkage".

	A	B	C	D	E	F	G	
A	0	0.5	0.43	1	0.25	0.63	0.38	
B	0.5	0	0.71	0.83	0.67	0.20	0.78	0.63
C	0.43	0.71	0	1	0.43	0.67	0.33	0.20
D	1	0.83	1	0	1	0.80	0.86	0.71
E	0.25	0.67	0.43	1	0	0.78	0.38	0.83
F	0.63	0.20	0.67	0.80	0.78	0	0.75	0.78
G	0.38	0.78	0.33	0.86	0.38	0.75	0	0.20

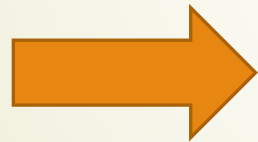


	BF
	0.63
	0.20
	0.71
	0.83
	0.78
	0.20
	0.78

Herarchical Clustering

- Se crea una nueva matriz de similitud

BF
0.63
0.20
0.71
0.83
0.78
0.20
0.78



	A	BF	C	D	E	G
A	0	0.63	0.43	1	0.25	0.38
BF	0.63	0	0.71	0.83	0.78	0.78
C	0.43	0.71	0	1	0.43	0.33
D	1	0.83	1	0	1	0.86
E	0.25	0.78	0.43	1	0	0.38
G	0.38	0.78	0.33	0.86	0.38	0

Herarchical Clustering

- El algoritmo se repite hasta que se realicen todos los clústers posibles.

	A	BF	C	D	E	G
A	0	0.63	0.43	1	0.25	0.38
BF	0.63	0	0.71	0.83	0.78	0.78
C	0.43	0.71	0	1	0.43	0.33
D	1	0.83	1	0	1	0.86
E	0.25	0.78	0.43	1	0	0.38
G	0.38	0.78	0.33	0.86	0.38	0

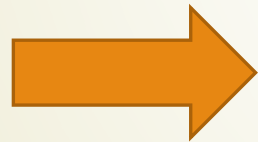


AE
0.25
0.78
0.43
1
0.25
0.38

Hierarchical Clustering

- Se crea una nueva matriz de similitud

AE
0.25
0.78
0.43
1
0.25
0.38



	AE	BF	C	D	G
AE	0	0.78	0.43	1	0.38
BF	0.78	0	0.71	0.83	0.78
C	0.43	0.71	0	1	0.33
D	1	0.83	1	0	0.86
G	0.38	0.78	0.33	0.86	0

Herarchical Clustering

	AE	BF	C	D	G
AE	0	0.78	0.43	1	0.38
BF	0.78	0	0.71	0.83	0.78
C	0.43	0.71	0	1	0.33
D	1	0.83	1	0	0.86
G	0.38	0.78	0.33	0.86	0



CG
0.43
0.78
0.33
1
0.33

Herarchical Clustering

CG
0.43
0.78
0.33
1
0.33



	AE	BF	CG	D
AE	0	0.78	0.43	1
BF	0.78	0	0.78	0.83
CG	0.43	0.78	0	1
D	1	0.83	1	0

Hierarchical Clustering

	AE	BF	CG	D
AE	0	0.78	0.43	1
BF	0.78	0	0.78	0.83
CG	0.43	0.78	0	1
D	1	0.83	1	0



AECG	0.43
0.78	
AECG	0.43
1	

AECG	0.43
0.78	
AECG	0.43
1	



	AECG	BF	D
AECG	0	0.78	1
BF	0.78	0	0.83
D	1	0.83	0

Herarchical Clustering

	AECG	BF	D
AECG	0	0.78	1
BF	0.78	0	0.83
D	1	0.83	0



AECGBF
0.78
0.78
1

AECGBF
0.78
0.78
1

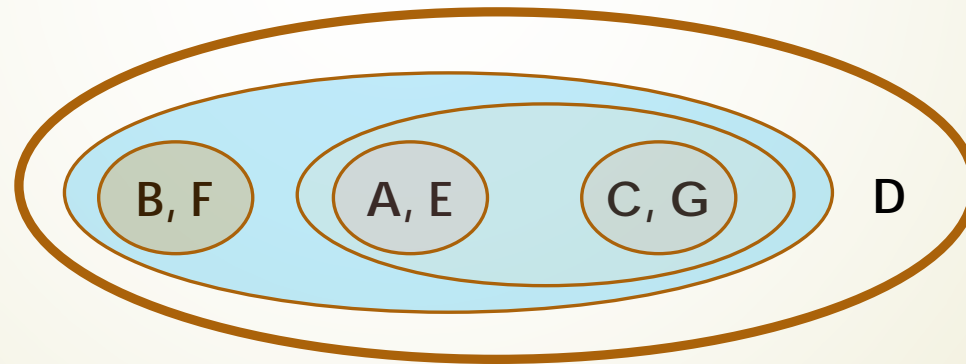


	AECGBF	D
AECGBF	0	1
D	1	0

Herarchical Clustering


- La interpretación de los resultados se puede realizar por medio de diagramas de Benn

A, B, C, D, E, F, G

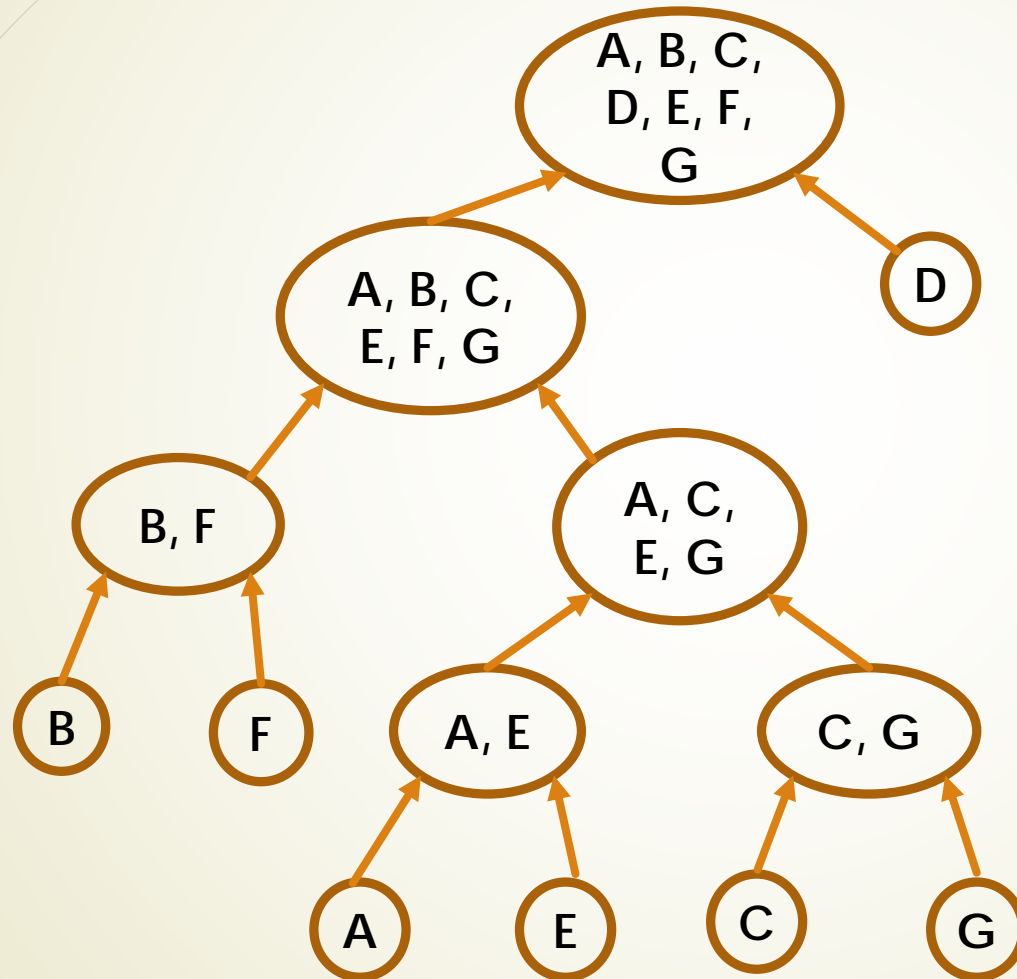




Herarchical Clustering

- ▶ La interpretación puede ser realizada también por medio de dendogramas.
 - ▶ Son estructuras con forma de árbol que sirven para visualizar agrupaciones o relaciones entre variables.
- 

Herarchical Clustering





Métodos de Aglomeración

Métodos de Aglomeración

➤ Complete

$$d(u, v) = \max(d(u[i], k[i]), d(v[i], k[i]))$$

➤ Single

$$d(u, v) = \min(d(u[i], k[i]), d(v[i], k[i]))$$

➤ Average

$$d(u, v) = \frac{n_u d(u[i], k[i]) + n_v d(v[i], k[i])}{n_u + n_v}$$

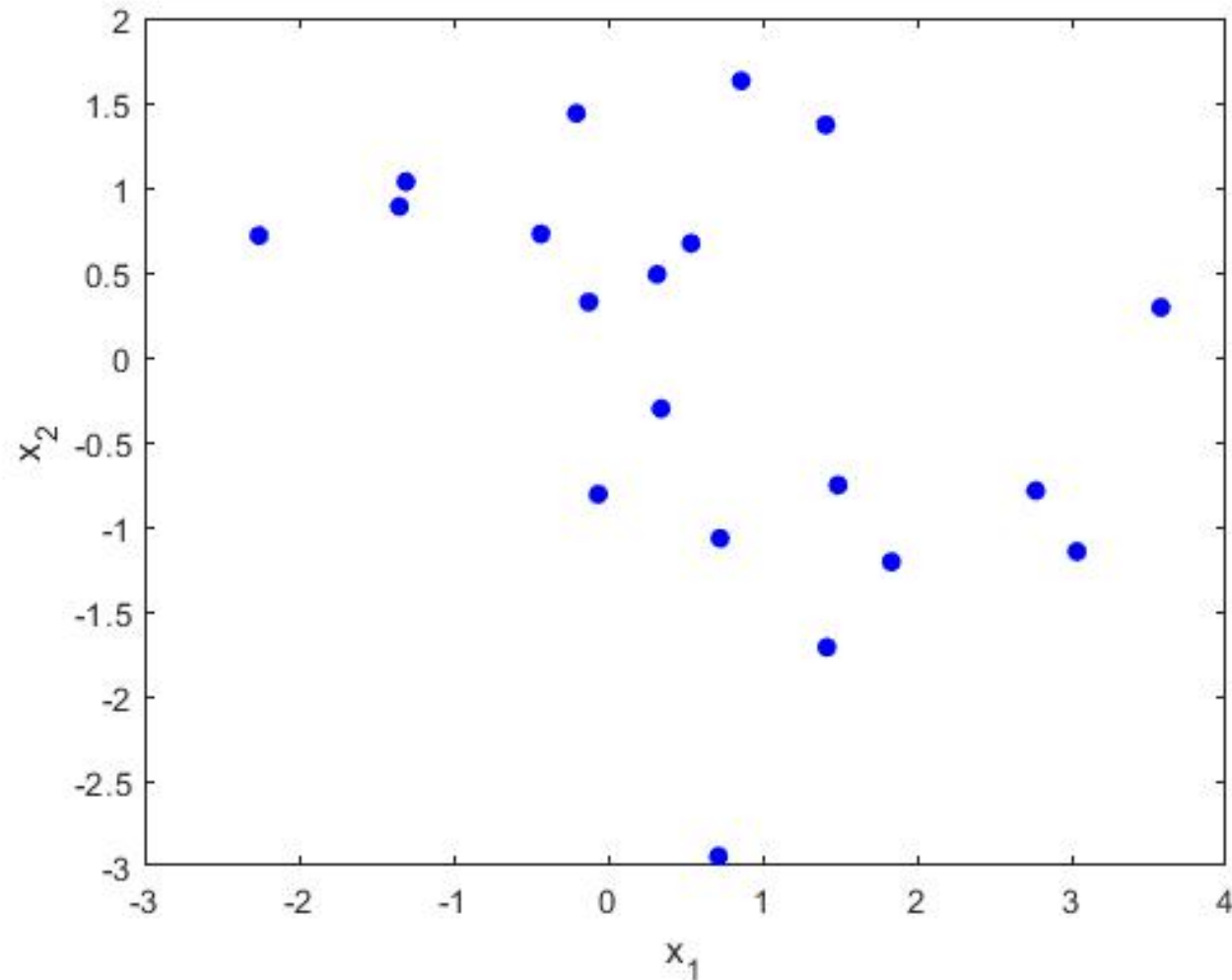
➤ Centroid

$$d(u, v) = \|c_u - c_v\|_2 = \sqrt{\frac{n_u d(u[i], k[i]) + n_v d(v[i], k[i])}{n_u + n_v} - \frac{n_u n_v d(u[i], v[i])}{(n_u + n_v)^2}}$$

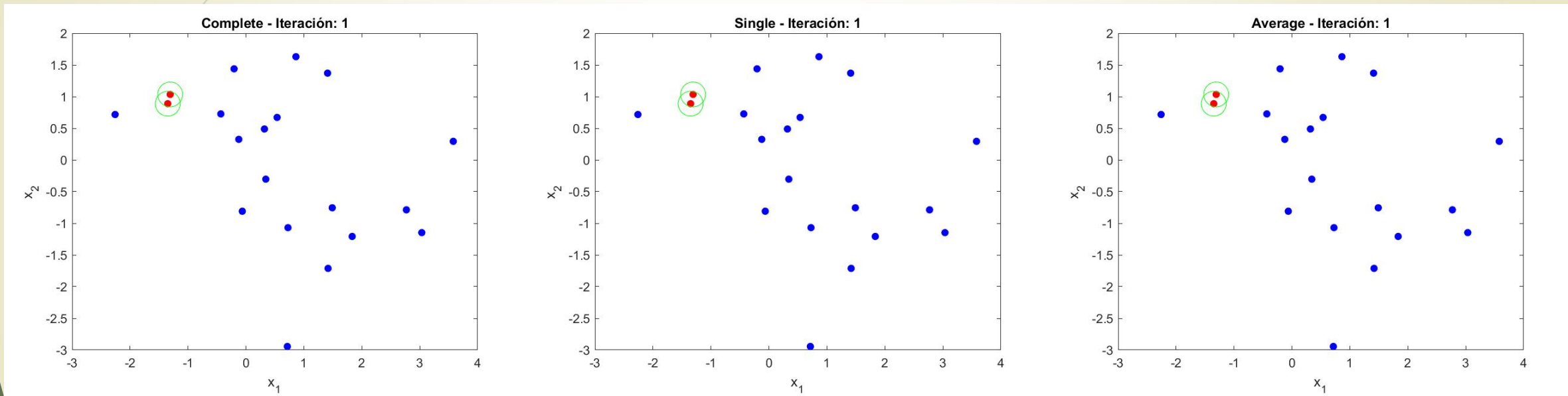
➤ Ward

$$d(u, v) = \sqrt{\frac{(n_u + n_k) d(u[i], k[i]) + (n_v + n_k) d(v[i], k[i]) - n_k d(u[i], v[i])}{n_u + n_v + n_k}}$$

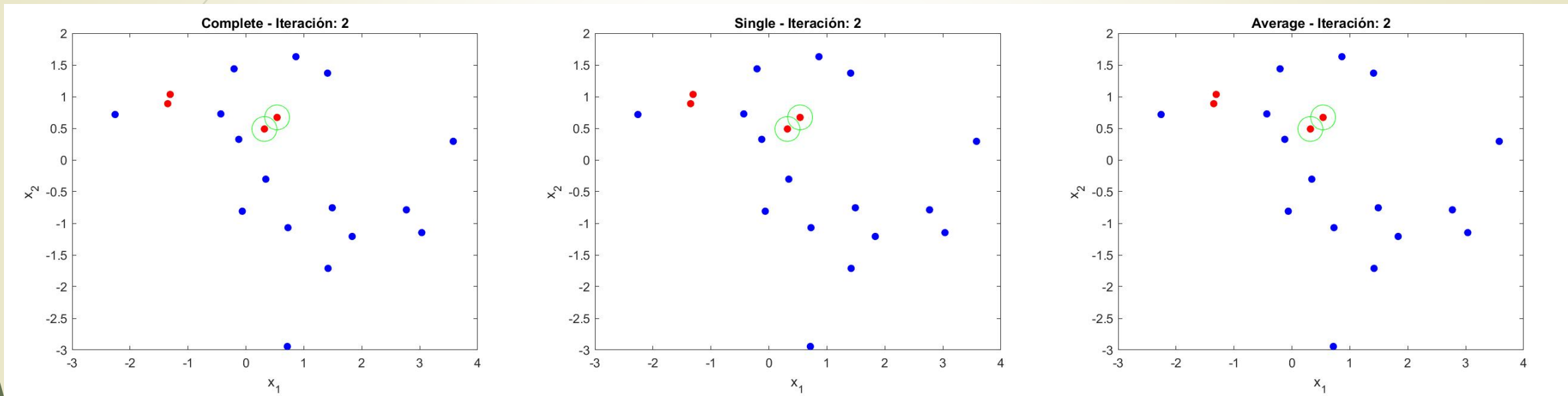
Complete vs Single vs Average



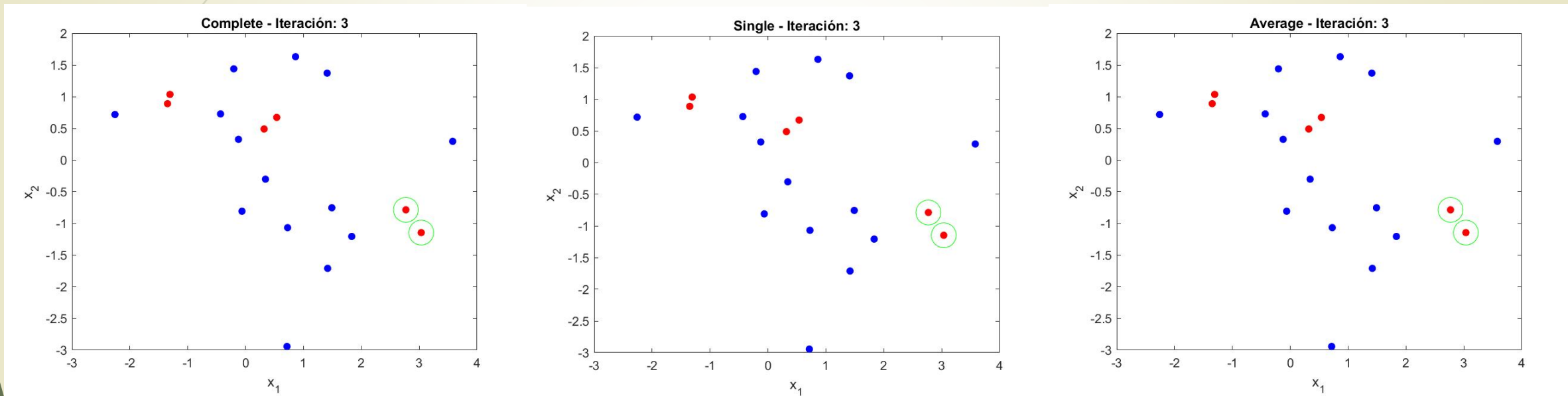
Complete vs Single vs Average



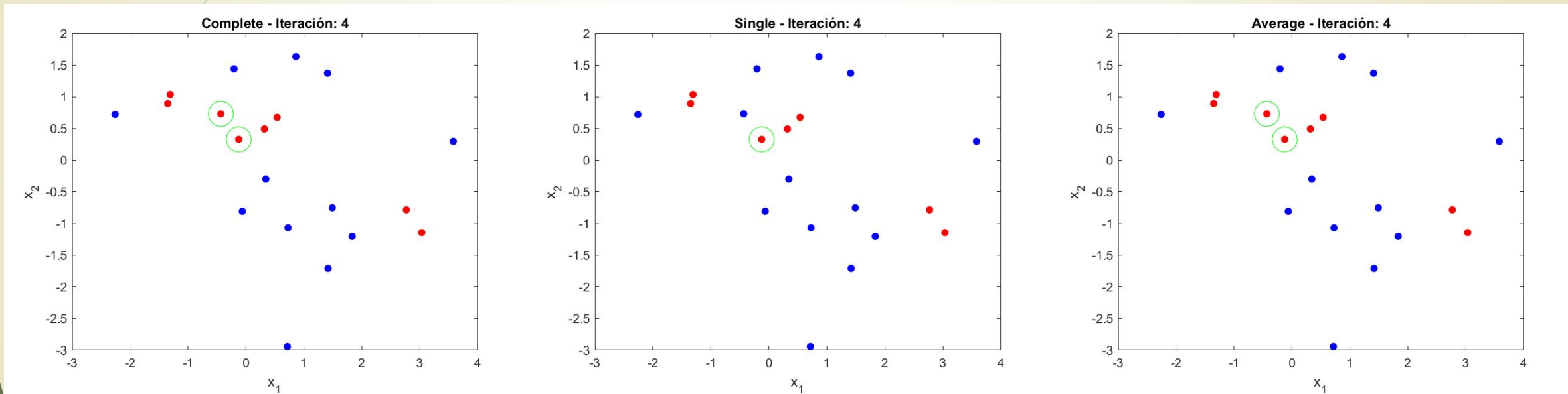
Complete vs Single vs Average



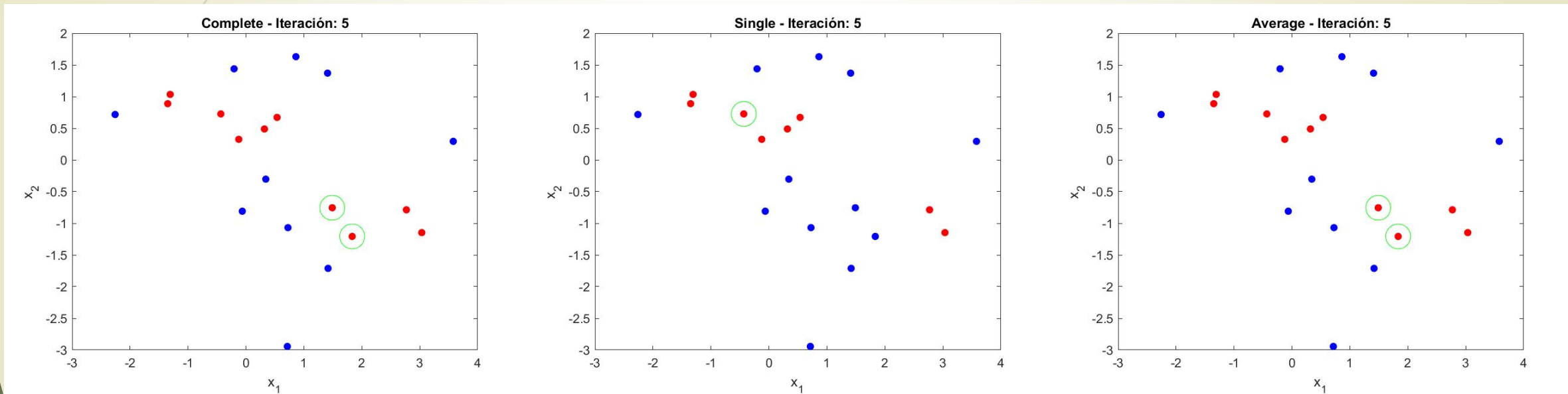
Complete vs Single vs Average



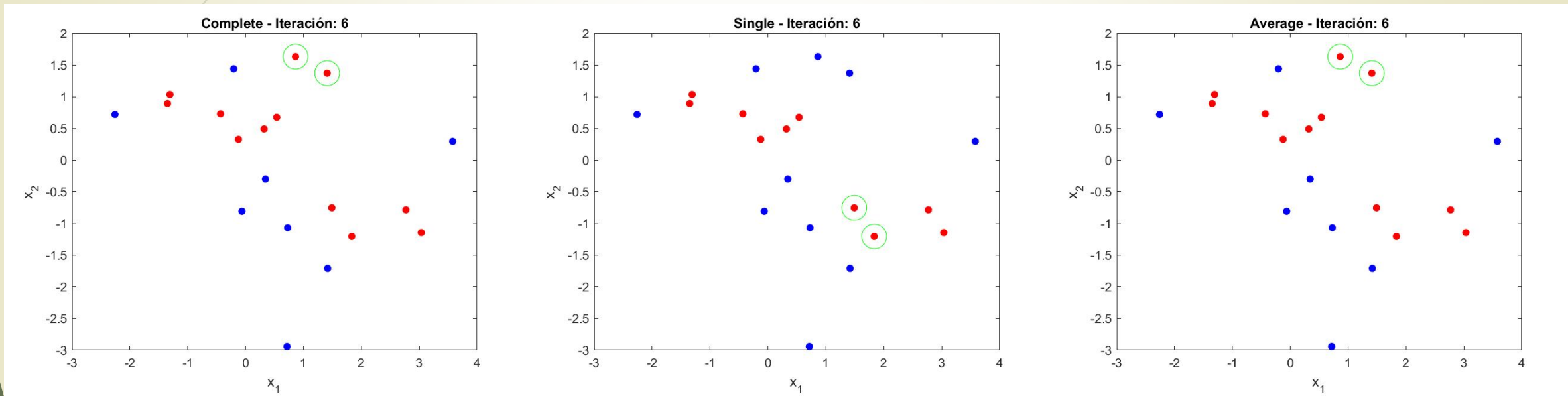
Complete vs Single vs Average



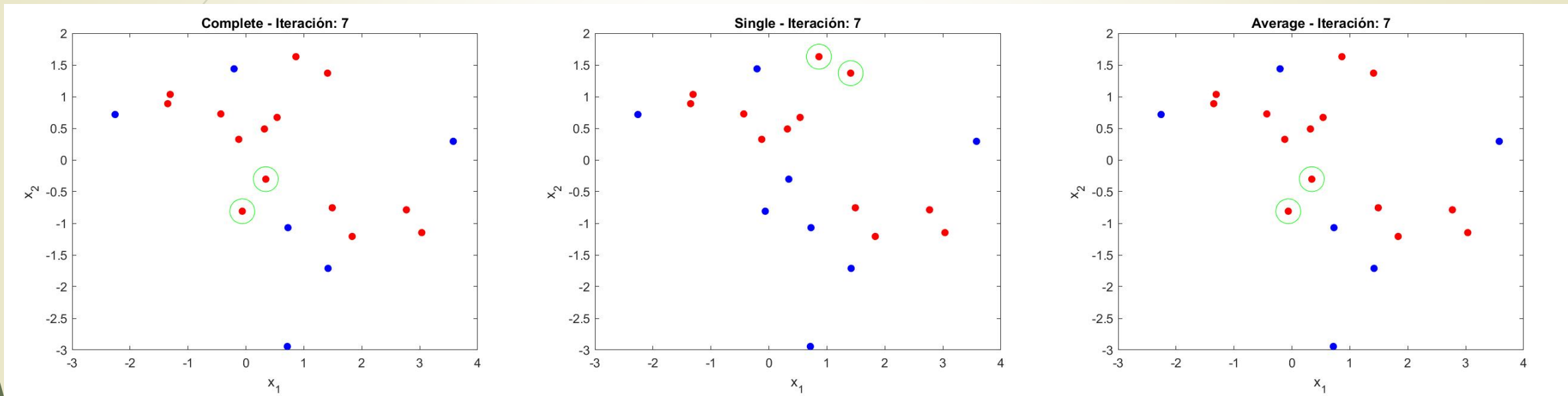
Complete vs Single vs Average



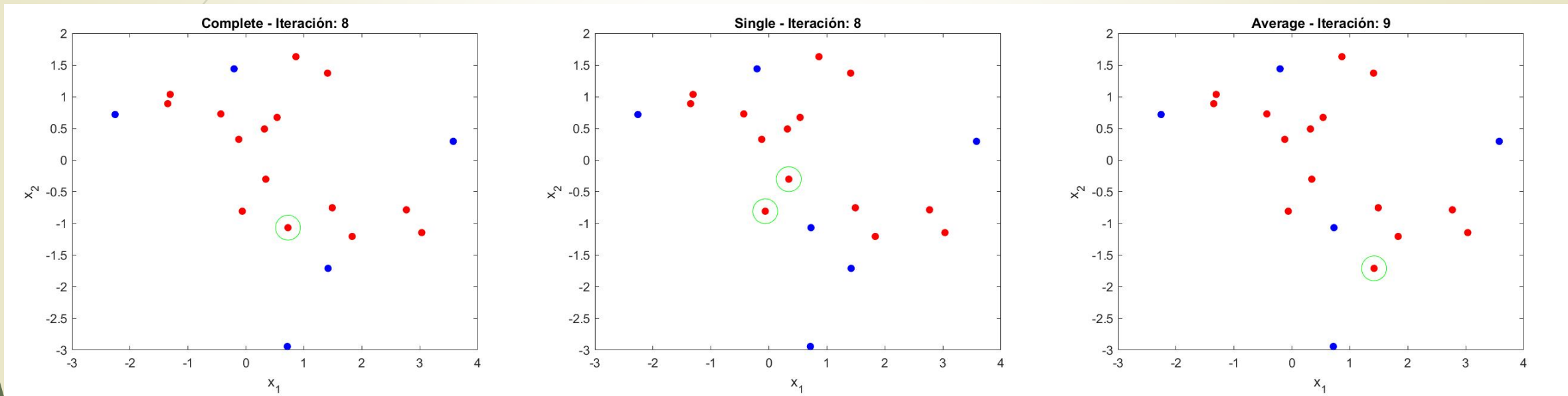
Complete vs Single vs Average



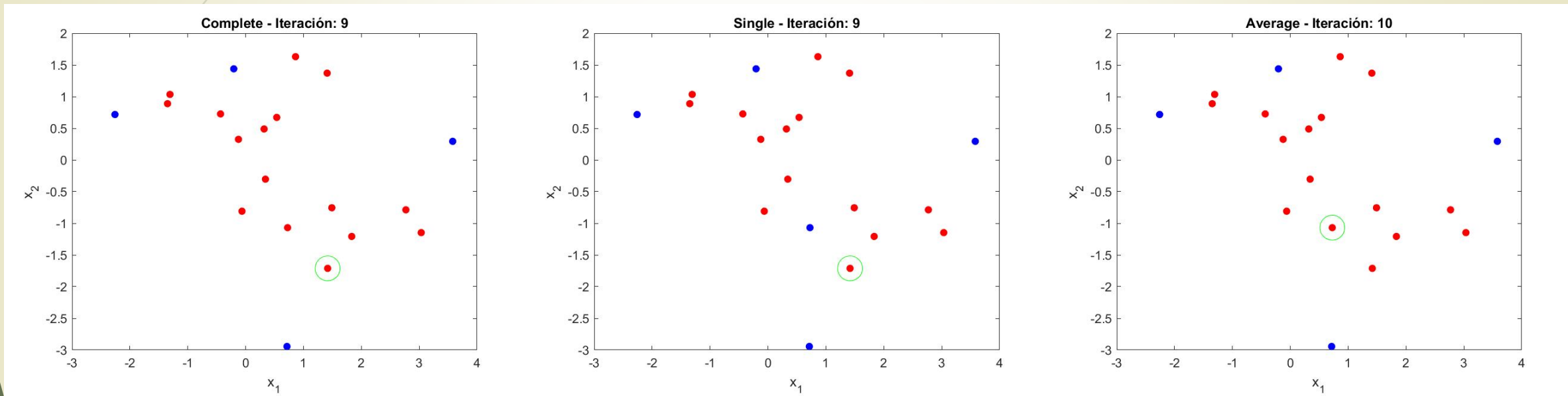
Complete vs Single vs Average



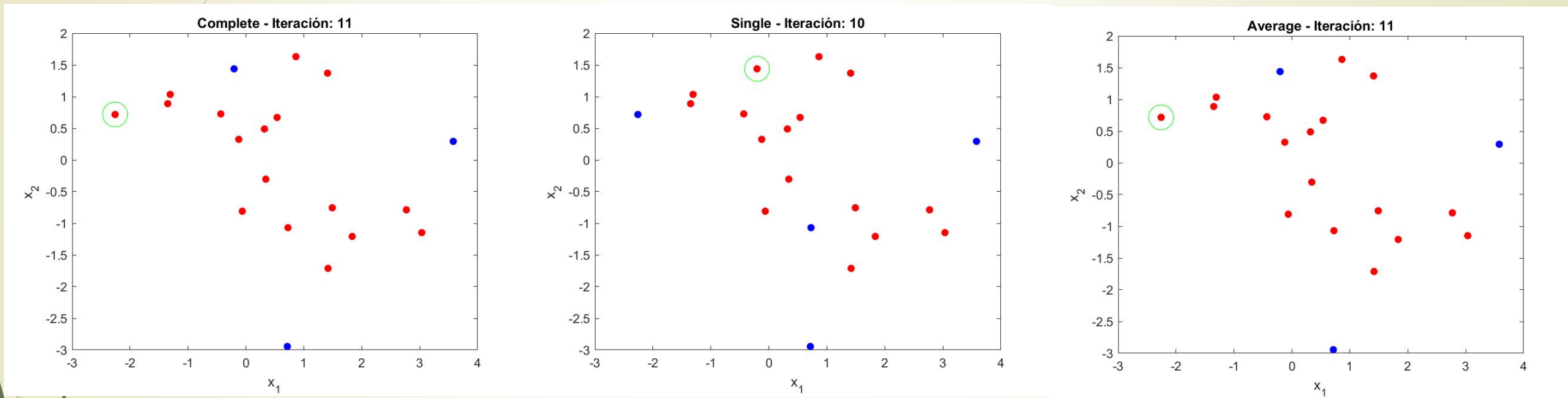
Complete vs Single vs Average



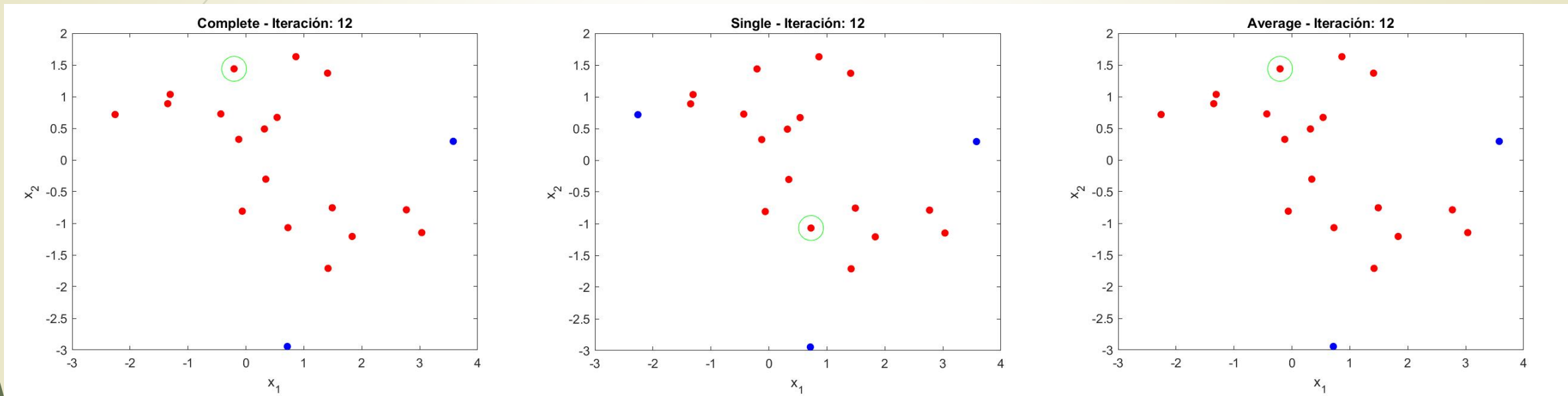
Complete vs Single vs Average



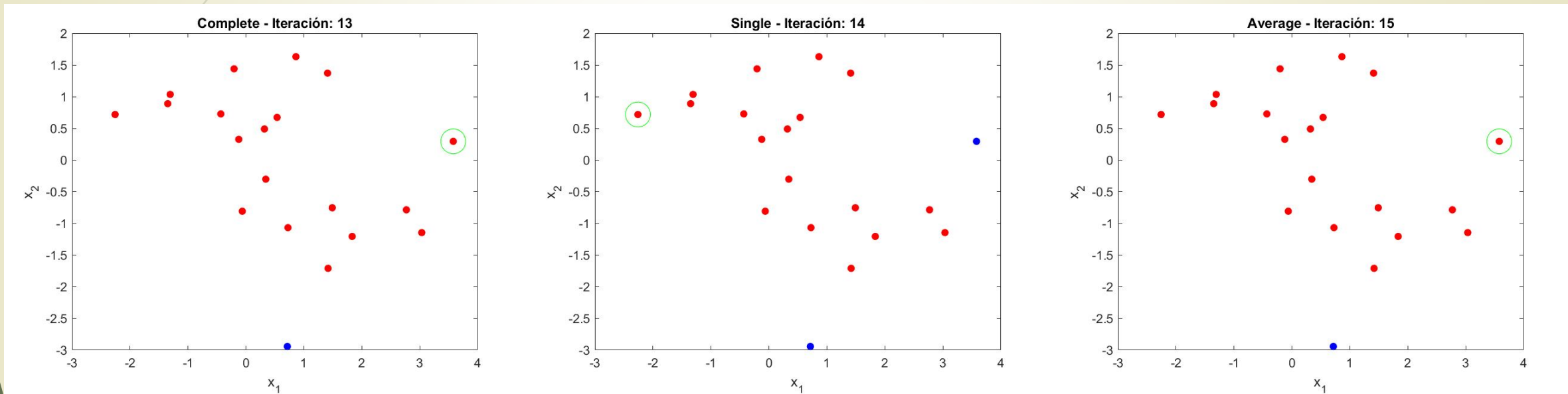
Complete vs Single vs Average



Complete vs Single vs Average



Complete vs Single vs Average



Complete vs Single vs Average

